TopicWeave: Enhancing Scientific Literature Topic Modeling through Distilled Model Ensembles - Milestone Report

Quddus Chong
MIDS, UC Berkeley
quddus.chong@berkeley.edu

Instructor: Jennifer Zhu
MIDS, UC Berkeley
jennifer.zhu@berkeley.edu

Abstract

This milestone report details the initial development of TopicWeave, a novel approach for scientific literature topic modeling employing distilled model ensembles. We describe our data acquisition and preprocessing of a balanced dataset comprising 50,000 arXiv papers across 10 scientific categories. We implement a baseline BERTopic model, using a general-purpose embedding model. Our baseline model demonstrates moderate semantic coherence but minimal alignment with ground truth categories, highlighting the limitations of general-purpose embeddings for scientific topic modeling. This milestone highlights the challenges of general-purpose embeddings in scientific topic modeling, setting the stage for our ensemble approach aimed at producing more coherent, interpretable, and scientifically relevant topics.

1 Introduction

The rapid expansion of scientific literature necessitates efficient methods for organization and insight extraction. Topic modeling offers a valuable approach to uncover latent thematic structures, aiding in tracking trends and reviewing literature. These structures also enhance information extraction and named entity recognition, improving the ability to identify entities and relationships within scientific text. Neural topic modeling approaches like BERTopic have shown effectiveness in general applications but often fail to capture domain-specific nuances in scientific literature due to its specialized vocabulary and frameworks. TopicWeave addresses this by proposing an ensemble approach that combines distilled versions of SciBERT and a citation-aware model, inspired by SPECTER2, to create a unified input representation for BERTopic. By leveraging both SciBERT's domain-specific understanding and citation-based relational knowledge, TopicWeave aims to produce more coherent, interpretable, and scientifically relevant topics while maintaining computational efficiency through model distillation.

2 Background

Neural topic modeling has evolved significantly from traditional probabilistic methods like Latent Dirichlet Allocation (LDA). (Abdelrazek et al., 2023) provide a comprehensive survey of topic modeling algorithms and applications, highlighting the shift toward neural approaches. BERTopic (Grootendorst, 2022) represents a significant advancement by leveraging neural embeddings to cluster documents and extract topics using TF-IDF, demonstrating superior performance compared to traditional methods.

Recent work has recognized the value of domain-specific adaptations for scientific literature. (Chagnon et al., 2024) introduced BERTeley, which streamlines topic modeling for scientific literature. However, as noted by (Kang et al., 2023), there remains significant potential for improving topic modeling through specialized language models. (Wolff et al., 2024) demonstrated that enriched BERT embeddings can enhance scholarly publication classification, providing evidence that domain-specific pre-training yields benefits for scientific document understanding.

The potential of ensemble approaches has been explored by (Voskergian et al., 2024), who showed that combining multiple topic models can improve text classification performance. However, their approach focuses on post-hoc ensemble methods rather than the integration of complementary embedding models at the representation level as proposed in TopicWeave.

Distillation techniques offer a promising avenue for maintaining the benefits of large pretrained models while reducing computational requirements. The SciBERT model (Beltagy et al., 2019) was specifically pre-trained on a large corpus of scientific text and has demonstrated superior performance on domain-specific tasks. Similarly, SPECTER2 ((Singh et al., 2022) captures inter-document relationships through training on citation graphs, providing complementary information about document relationships that pure text-based models may miss.

Recent work by (Mu et al., 2024) suggests that large language models can offer alternative approaches to traditional topic modeling, indicating a broader trend toward leveraging advanced language models for this task. (Wijanto et al., 2024) explored optimal hyperparameter tuning for BERT-based topic models applied to scientific articles, highlighting the importance of careful parameter selection for these approaches.

TopicWeave builds upon these advances by combining domain-specific embeddings through an ensemble approach while ensuring computational efficiency through model distillation. The key innovation in our neural architecture is the knowledge-specialized dual pathway approach, where two parallel embedding models capture different aspects of scientific understanding: domain-specific vocabulary (SciBERT) and document relationships (citation-aware model).

3 Methods

3.1 Dataset Acquisition and Preprocessing

We extracted a balanced subset of 50,000 papers from the arXiv dataset (Cornell University, 2024) across 10 scientific categories, using author-assigned categories as external validation metrics. The preprocessing pipeline retained document IDs, titles, abstracts, and primary categories, with text cleaning involving stopword removal and lemmatization. Titles and abstracts are then concatenated to create single document representations.

3.2 Baseline Model Implementation

A baseline BERTopic model was implemented using the *paraphrase-MiniLM-L12-v2* sentence transformer as the embedding model, without domain-specific pre-training, configured with a minimum topic size of 10 documents. We configured the model with a minimum topic size of 10 documents to ensure meaningful topic representations. The baseline followed BERTopic's standard pipeline: embedding generation, dimensionality reduction using UMAP, clustering with HDBSCAN, and topic extraction using c-TF-IDF.

3.3 TopicWeave Architecture Implementation

The TopicWeave architecture involves a multi-stage approach: we first preprocess input documents and create two distilled domain-specific models: a distilled SciBERT model and a citation-aware model. The distilled SciBERT model was created by pruning the number of transformer layers from 12 to 6 while maintaining the original vocabulary and embeddings. This reduced model is then trained to mimic the full SciBERT model's sentence embeddings using mean squared error loss. The citation-aware model, also with a 6-layer architecture, was trained using a contrastive learning approach, inspired by (Gao et al., 2021), to recognize relationships between sections of the same document versus different documents.

Each model independently generates embeddings for all preprocessed documents, capturing complementary aspects of scientific text. The individual embeddings are then combined using an optimized weighted average to create a unified representation. Weights are optimized based on silhouette score, NMI, ARI, and within-category similarity.

We implemented this distillation pipeline and embedding generation process using Python and PyTorch, using the Hugging Face Transformers and SentenceTransformer libraries. It runs efficiently on Google Colab, requiring no specialized hardware.

3.4 Evaluation Metrics

Evaluation metrics include Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), C_v coherence, and Normalized Pointwise Mutual Information (NPMI).

We evaluate both the baseline and TopicWeave models using quantitative metrics (NMI between discovered topics and ground truth categories, ARI to assess clustering similarity) and qualitative metrics (Topic coherence using C_v measure, NPMI to assess semantic relatedness of topic terms).

4 Results and Discussion

4.1 Baseline Model Performance

The baseline BERTopic implementation, using a general-purpose embedding model, produced 256 topics from the arXiv dataset. Quantitative evaluation revealed moderate topic coherence, as indicated by a C_v score of 0.5486, but limited term co-occurrence (NPMI: 0.0184). Alignment with

Table 1: Baseline BERTopic Performance on arXiv Dataset

Metric	Value
NMI	0.2460
ARI	0.0017
C_v Coherence	0.5486
NPMI Coherence	0.0184
Topics (excl. outliers)	256

ground truth categories was weak, with a Normalized Mutual Information (NMI) of 0.2460 and an Adjusted Rand Index (ARI) of 0.0017, suggesting minimal agreement beyond chance. These results are summarized in Table 1. The model generated 256 topics from 22,699 documents, indicating finer-grained distinctions than the broad categories, though these distinctions did not align well with expert-defined categories.

Analysis of top topics revealed meaningful themes but also a substantial outlier topic to which nearly half the dataset (11,068 documents) is assigned. As shown in Table 2, term weight distributions followed a power-law pattern, with top terms having higher weights, and weights converging around the 6th-8th ranked term.

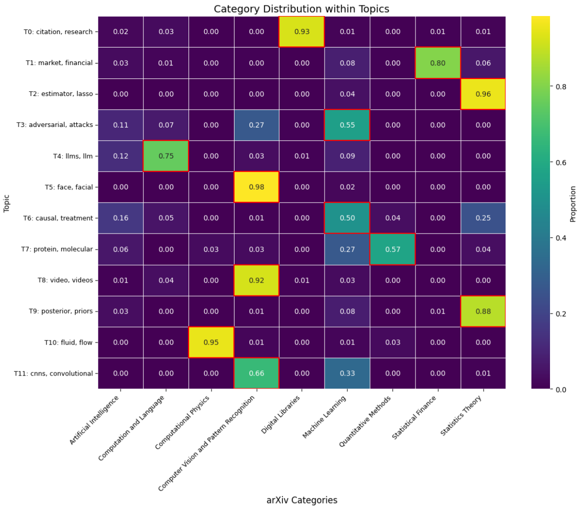

Figure 1: Topic-category alignment, revealing both category-specific topics and interdisciplinary themes spanning multiple scientific domains.

These observations reveal limitations in the baseline model's ability to identify consistently distinctive terminology across different scientific domains and capture the specialized vocabulary of scientific discourse.

Table 2: Top Topics from Baseline Model

ID	Representative Terms	Primary Category
0	citation, research	Digital Lib. (93%)
1	market, financial	Stat. Finance (80%)
2	estimator, lasso	Stat. Theory (96%)
3	adversarial, attacks	ML (55%)
4	llms, chatgpt	Comp. Lang. (75%)

5 Next Steps

Having completed the implementation and execution of our distillation and embedding generation pipeline in Google Colab, our immediate next steps focus on comprehensive evaluation and refinement of the approach.

Future steps include integrating TopicWeave embeddings with the BERTopic pipeline and conducting direct comparisons with the baseline, optimizing BERTopic hyperparameters, and performing ablation studies with individual embedding types. We aim to conduct deeper qualitative analyses of topic interpretability, coherence, hierarchies, and distinctiveness across domains, along with performance benchmarking for computational efficiency. Future extensions could include improving the citation model with Graph Neural Networks and citation context analysis, enhancing the embedding combination mechanism with learnable combinations, and exploring dynamic topic modeling with citation-awareness. Additional evaluation metrics for novelty detection and implementation enhancements for scalability and hyperparameter optimization will be considered.

These steps aim to demonstrate that TopicWeave's ensemble approach with distilled domain-specific models provides meaningful improvements in scientific literature topic modeling while maintaining computational efficiency.

References

Aly Abdelrazek et al. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Eric Chagnon et al. 2024. Benchmarking topic models on scientific articles using berteley. *Natural Language Processing Journal*, 6:100044.

Cornell University. 2024. arxiv. Kaggle, 27 Feb. 2024, www.kaggle.com/datasets/Cornell-University/arxiv.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Woojin Kang et al. 2023. An analysis of research trends on language model using bertopic. In *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)*. IEEE.

Yida Mu et al. 2024. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.

Amanpreet Singh et al. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *arXiv preprint arXiv:2211.13308*.

Daniel Voskergian, Rashid Jayousi, and Malik Yousef. 2024. Topic selection for text classification using ensemble topic modeling with grouping, scoring, and modeling approach. *Scientific Reports*, 14(1):23516.

Maresha Caroline Wijanto, Ika Widiastuti, and Hwan-Seung Yong. 2024. Topic modeling for scientific articles: Exploring optimal hyperparameter tuning in bert. *International Journal on Advanced Science, Engineering & Information Technology*, 14(3).

Benjamin Wolff, Eva Seidlmayer, and Konrad U Förstner. 2024. Enriched bert embeddings for scholarly publication classification. In *International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs*. Springer Nature Switzerland.